Audience Reach Projection

for the 2010 FIFA World Cup in South Africa

Valeria Montero Garnier

Wharton Research Scholars

May 10, 2010

My project with the Wharton Interactive Media Initiative (WIMI) aims to model individual consumption behavior of 2010 World Cup content across ESPN's media platforms. We specified and tested two Hidden Markov models on simulated data to forecast mutli-platform reach. Currently we are developing a Bayesian multivariate regression with time-varying covariates. Communicating with ESPN executives and analyzing data from the 2010 NCAA March Madness Basketball championship allowed us to identify expected reach patterns for digital platforms. The March Madness dataset also helped us understand how the World Cup data will be measured. Our next steps are to improve the specifications of the Bayesian regression model to better accommodate the correlations, heterogeneity, and non-stationarity of the expected data. Once the first two weeks of the World Cup data become available, we will apply and refine our models to forecast the digital cross-platform reach during the last three weeks of the championship. A predictive model from multi-platform reach will help ESPN better serve customers and more accurately value advertisements.

Objectives

ESPN provides sports content and sells advertisements through many platforms including Internet (ESPN.com), streaming video (ESPN360 broadband service to more than 50 million suscribers), mobile phones (ESPN Mobile Properties provides live game coverage, news, and alerts), radio (ESPN Radio Network + 5 radio stations), print, and television. The WIMI project is part of ESPN's cross-platform research initiative for the 2010 FIFA World Cup in South Africa. The initiative, ESPN XP, aims to research behavioral patterns across all media platforms around the upcoming World Cup. WIMI's goal is to understand the drivers of consumer

behavior across digital platforms (Internet, video, and mobile) and to forecast multichannel consumption for the last three weeks of the championship. Our model will be calibrated on the group matches between groups A-H during June 11 to June 25, 2010 and predict reach for the Round of 16, Quarter Finals, Semi Finals, $\frac{3}{4}$ Place, and Final phases of the World Cup (June 26 – July 11). Platform-specific and cross-platform predictions will incorporate covariates to understand the drivers of digital platform usage, such as weekend versus weekend effects. Our efforts will be complemented with parallel projects by Nielsen, Knowledge Networks, the Media Behavior Institute, and the Keller Fay Group. These partners will collect consumption data including away-from-home viewership and conduct parallel projects such as analyzing word of mouth trends around the championship sponsors. Our combined efforts will provide key insights to better value and allocate advertisements for future events. Glenn Enoch, the Vice President of Integrated Media Research at ESPN, describes how the initiative as one that will revolutionize the industry: "This is about the future of cross-media measurement. We are working with these companies to help develop a model for the industry, to advance knowledge about cross-media research and behavior, to find techniques that work and discard ones that don't. We want to bring work in this area closer to currency measures and bring the industry closer to a day when measuring cross-platform behavior is a standard practice instead of a special project"[6].

The data for the Internet platform is collected from a panel of more than 10,000 people. The data for video streaming is retrieved from Nielsen's Life 360 mobile platform. Our analysis focuses on the only 13% of ESPN's total customers because relatively few people are registered for the mobile application star wave ID's (SWIDS). A similar data set from the NCAA March Madness 2010 basketball championship guided our model specifications by providing insights about underlying behaviors. For instance, we decided to focus on modeling reach rather than frequency because the March Madness dataset revealed that Gamecast, a program which automatically refreshes web pages, artificially increased page view counts.

Three Modeling Challenges

Models for reach across platforms during the World Cup integrate correlations of cross-platform viewership, non-stationarity, and consumer heterogeneity.

Correlation in Behavior across platforms

Correlations across platforms illustrate trends in cross-platform usage. Consider a segment which often checks news on the mobile service while at work but does not have time to stream videos. On the weekends however, they spend most of their time surfing the Internet and streaming videos instead of using the mobile service. Taking into account correlations into our model avoids double counting people across platforms. For example, if we observe 20 visits to espn.com and 20 visits to streaming videos, and we know that there

is a perfect positive correlation between both platforms, then we can infer a reach only 20 customers rather than 40.

Copulas can model non-linear correlations between consumption patterns for each platform. A copula is a multivariate distribution function with uniform marginal distributions. Sklar's Theorem allows us to model dependence through copulas: For random variables $x_{1,...,}x_n$ with marginal distributions $F_{1,...,}F_n$, a unique copula C relates the marginal distributions F_i and the multivariate distribution F: $F(x_{1,...,}x_n) = C(F_1(x_1), ..., F_n(x_n))$. In other words, a joint distribution can be decomposed into the marginal distributions of each random variable and a copula that specifies the dependency between the random variables. Rearranging the equation above to solve for C: $C = F(F_1^{-1}(y_1), ..., F_n^{-1}(y_n))$. Many types of copulas exist. C is called a Gaussian copula when F is a multivariate standard normal distribution. C is called a t-copula when F is a multivariate t distribution. The t-copula is similar to the Gaussian copula because the relationship between the marginals is symmetric. The t-copula generates larger correlation for large co-movements because the tails of the marginal distributions have a higher dependence. Danaher and Smith have explored these methods in Modeling Multivariate Distributions Using Copulas: Applications in Marketing (Forthcoming in Marketing Science, 2009). Their paper implements copulas to model situations including advertising exposure to magazine advertisements and page views across websites. Other publications have used the Sarmanov distribution, which can be represented as a specific type of copula, for similar applications such as the bacon and eggs example (Modeling Browsing Behavior at Multiples Websites, Park and Fader, 2004).

Hidden Markov models and regressions are alternative ways we pursued to model correlations across platforms. The advantage of using regression is that it can very easily accommodate many time-varying covariates.

Non-stationarity and consumer heterogeneity

A stochastic process is non-stationary when the probability distribution of the process changes over time. Imagine a US fan who frequently consumes digital content on digital platforms at the start of the World Cup, then loses interest when the US fails, and finally begins to avidly watch television content during the final, but does not regain interest on the digital platforms. The variation in viewership propensities across platforms over time is difficult to model because they are strongly affected by the outcomes of the championship phases. Consumer heterogeneity means that although some consumers behave similarly, everyone's consumption propensities vary. The consumption patterns of a 50 year old manager will be very different to those of a college student.

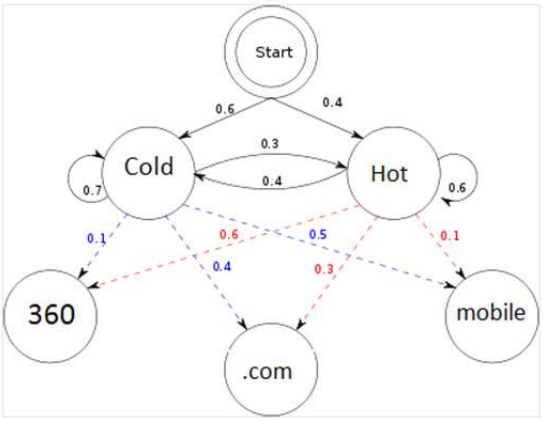
Covariates can capture some of the non-stationarity and consumer heterogeneity. Observable covariates specific to the panelists, such as demographics, can capture some of the differences in intrinsic interests between consumers but not all of them. Tournament phases, weekday/weekend effects, and whether specific

teams play each day or not are time-varying covariates that can track some of the non-stationarity in viewership behavior. Instead of looking at whether a specific team plays each day, we will calculate team-attraction parameters during the initial championship phase when groups A through H play against each other. The estimates can then be used to forecast reach under different scenarios. Each of the initial 32 teams will play 3 times during the first phase. If three games per team are not enough to estimate reliable attraction parameters then we will estimate group-specific parameters. IRT (Item Response Theory) is a model for team-specific attraction parameters where the likelihood of incidence (view/day) is a function of whether each team played on a specific day or not. Let θ_k be the attraction parameter of team k. Let I_t be an indicator variable of 1 if team k had played on day t. The reach for a given day will be modeled as a function of $\sum_{k=1}^{K} \theta_k I_t$. ESPN managers expect that the US and Latin American teams will be the most popular.

Three Model Specifications

HIDDEN MARKOV MODELS (HMMs)

One way to model non-independent sequential data is to link each observation to the previous one. A first-order Markov model assumes that whether an individual frequently checks game scores on the Internet or not on a given day is affected by their state during the previous day. Our model envisions customers switching over time between a high propensity 'hot consumption' and a low propensity 'cold' state for each platform: $S \in [s^{cold}_{.com}, s^{hot}_{.com}, s^{cold}_{video}, s^{hot}_{video}, s^{cold}_{mobile}, s^{hot}_{mobile}]$. Markov models contain state transition probabilities that indicate the likelihood of staying in the current state or switching to a different state given the current state. In a hidden Markov process we cannot observe the true states of an individual but instead, we only observe the outcomes. For example, 50 daily page views is more likely to indicate a true 'hot' state instead of than 5 page views. The picture below adapted from an online illustration[7] shows how individuals switch between high and low states over time. The probability arrows between the cold and hot states symbolize the transition probabilities. Our models also allow for individuals to be in different states for different platforms. For example, an individual may experience a high propensity state for Internet and a low propensity state for mobile on the same day.



Bernoulli HMM

Our first model specifies a Bernoulli process to model daily platform reach. Let MVN represent the Multivariate Normal distribution, IW the Inverse Wishart distribution and I a diagonal matrix. S_{ijt}^{H} is an indicator variable that equals 1 if person i is in the high propensity state on platform j for day t, or equals 0 if the person is instead 'cold'. Let y_{ijt} (the observed outcome of the current state) be a binary observation of whether person i observed content on platform j on day t.

$$y_{ijt} \sim Bernoulli\left(rate = \frac{exp(\theta_{ijt}^{L}) + S_{ijt}^{H} * exp(\theta_{ijt}^{H})}{exp(\theta_{ijt}^{L}) + S_{ijt}^{H} * exp(\theta_{ijt}^{H}) + 1}\right)$$

Let $\bar{\theta} = [\theta_{i1t}^{L}, \theta_{i1t}^{H}, ..., \theta_{iJt}^{L}, \theta_{iJt}^{H}]$, where $\bar{\theta} \sim MVN(\mu_{2xJ}, \sum)$. We add conjugate prior distributions for μ and $\sum$:

$$\mu \sim MVN(0,\ I * C_m)$$

$$\sum \sim IW(\Psi = C_{\sum} * I,\ df)$$

$$Where\ \mu = [\mu_1^L, \mu_1^H, ..., \mu_J^L, \mu_J^H]$$

Derivation of the rate specification. Initially we envisioned a model with the following process:

$$y_{ijt} \sim Bernoulli\left(rate = \lambda_{ijt}^{Low} + S_{ijt}^{H} * \lambda_{ijt}^{H}\right)$$

One method to ensure a $rate = \lambda_{ijt}^{Low} + I_{ijt}^{H} * \lambda_{ijt}^{H}$ between zero and one is to define the low rate as a fraction

of the higher rate: $\lambda^L = \lambda^H * q$, where $q \in [0,1]$. Instead we use a logit to restrict rates between the range [0,1]. p^L is the probability of being in the low, or cold state. θ^L is the log odds of being in the cold state.

$$\theta^L = log\left(\frac{p^L}{1-p^L}\right)$$

$$\Rightarrow p^L = \frac{exp(\theta^L)}{exp(\theta^L)+1}$$

The probability of being in a hot state p^H guarantees $p^H > p^L$. Note that both probabilities are also restricted between zero and one.

$$p^H = \frac{exp(\theta^L)+exp(\theta^H)}{exp(\theta^L)+exp(\theta^H)+1}$$

The final expression uses an indicator variable for the hot state S^H_{ijt} to describe both $p^H$and p^L in one statement. The rate is equal to p^L when the indicator variable is zero. The rate is equal to p^H when the indicator variable is 1.

$$rate = \frac{exp(\theta^L_{ijt})+S^H_{ijt}*exp(\theta^H_{ijt})}{exp(\theta^L_{ijt})+S^H_{ijt}*exp(\theta^H_{ijt})+1}$$

Poisson HMM

We specified a model for daily consumption frequency instead of daily incidence. Although in similar settings the Poisson process might better forecast reach than the Bernoulli process, we prefer the Bernoulli HMM because our count data is biased by automatic page-refresher programs. The following is the specification of the Poisson HMM. We tested this model excluding the correlation parameter c and the non-stationarity parameter γ. Let y^p_i be the number of views of person i in platform p:

$$y^p_i \sim Poisson(rate = \lambda^p_i + \gamma_t)$$

Heterogeneity and correlation: Every person can be in a low or high state on each platform. The individual's λ rate is drawn from a platform-specific state-dependent Gamma distribution. Both the high and the low states for each platform have different a and b parameters. The c parameter, common among all of the high (low) state Gamma distributions, introduces correlation between platforms.

$$\lambda^p_i \sim \begin{cases} \Gamma(a^{l,p}, b^{l,p}+c^l) & if\ in\ low\ state \\ \Gamma(a^{h,p}, b^{h,p}+c^h) & if\ in\ high\ state \end{cases}$$

Non-stationarity is introduced through a time-varying gamma parameter that can be common across all platforms and states. Let X_t be the number of teams left in the cup, or a process with a drift.

$$\gamma_t \sim f(X_t)$$

Multivariate Logistic Regression

A multivariate logistic regression model has more flexibility to incorporate covariates than hidden Markov models. Although we specified an analogous frequency version, the incidence model for daily reach will help us manage the biased page view counts.

$$logit(P_{ijt}) = \mu_j + X\beta + e_{ijt} + \delta_{jt}$$

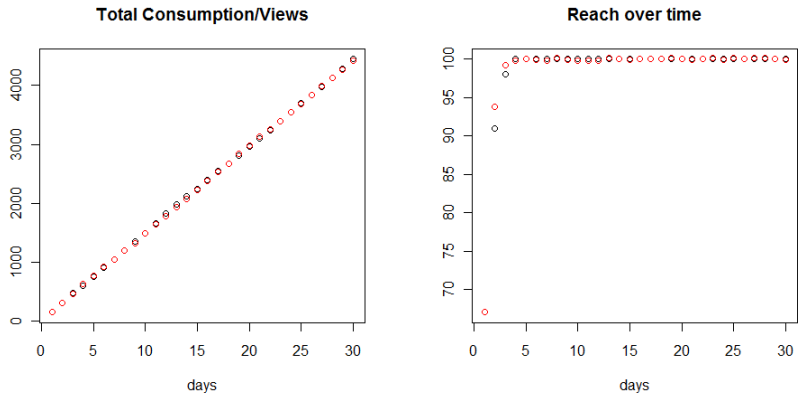
The first term, a mean incidence rate per platform, is drawn from a common multivariate normal distribution across platforms: $\mu_{1,...}\mu_K \sim MVN(0, 100 * I)$. The term $X\beta$ represents the covariates and their estimated β coefficients. e_{ijk} is an error term that is normally distributed with an Inverse Wishart prior distribution on the variance-covariance matrix. The last term δ_{kt} introduces a downward trend in digital platform consumption towards the end of the World Cup. The downward trend in digital consumption, observed in the 2010 NCAA March Madness data, was perhaps driven by an increased use of television during the final phases.

Estimation Results

We conducted posterior predictive tests and parameter recovery checks for the Bernoulli and Poisson hidden Markov models on simulated data. Posterior predictive tests compare the original data to data generated from a calibrated model. The first 10 out of 30 days of simulated data were first used to calibrate a version of the Poisson HMM without individual-level heterogeneity in the parameters. λs represent consumption propensities, ν represents the transition matrix, and init represents the initial state probabilities. We simulated a world with only 2 platforms.

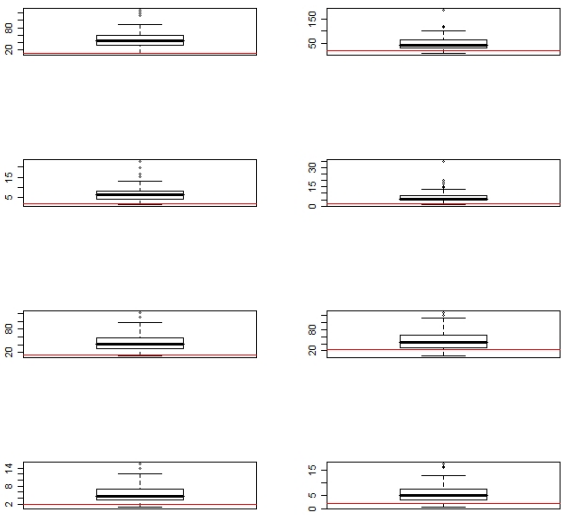
	True values	Parameters estimated first 10 days
λ_{mob}	(1,2)	(0.88, 1.97)
$\lambda_{.com}$	(2,4)	(2, 3.9)
ν	(0.7,0.3)(0.3,0.7)	(0.67, 0.33)(0.26, 0.74)
init	0.5	0.48

Total cumulative consumption is the sum of all of the views across platforms and individuals for each day. Reach is number of people who used at least 1 platform since day 1. The black dots represent the true cumulative reach and total consumption from 30 days of simulated data. The red dots represent daily reach and consumption for 30 days of data generated from the calibrated model. There is an artificial jitter in the reach plot to distinguish between overlapping red and black dots.

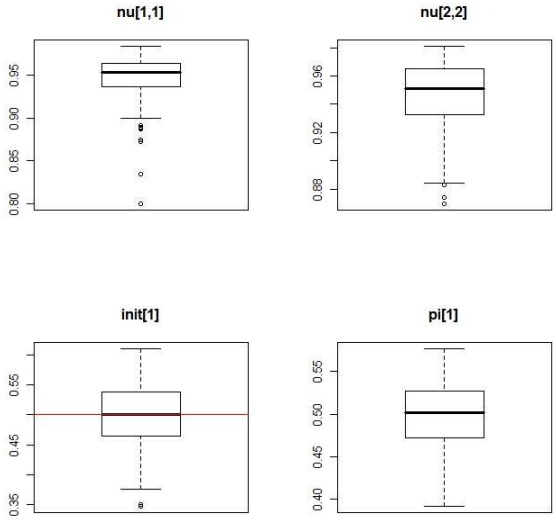


Adding heterogeneity to the Poisson HMM

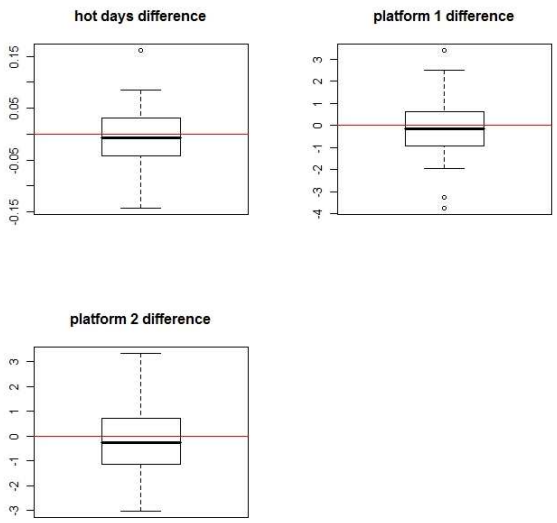
The previous model is modified to include heterogeneous lambda parameters across individuals. The correlation parameter c and the non-stationarity parameter γ are not included. The first figure below shows box plots of the a and b parameter estimates for the Gamma distributions. These replace the platform-specific lambda (high/low) parameters in the model without heterogeneity. The red line indicates the true values. The true Gamma parameters for platform 1 are $a = 10$, $b = 2$ for the low state and a=20, b=2 for the high state. The true gamma parameters for platform 2 are $a = 12$, $b = 2$ for the low state and $a = 24$, $b = 2$ for the high state. The first 4 illustrations represent the parameters for platform 1:



The following plots illustrate the parameter estimates for the transition matrix ν and the initial state probabilities. The top two plots show that the parameters for the transition matrix were overestimated (they show average estimates of around 0.95 when the true values are 0.7). The true initial state probability is 0.5.



Adding heterogeneity into the parameters yields poor estimates for the Gamma 'high' state and Gamma 'low' state distributions for each platform as well as for the common transition matrix. Despite the biased parameters, the resampled data from the calibrated model is similar to the original simulated data. The top left plot below shows the difference between the resampled and original average number of days in the 'high' state. The top right plot illustrates the difference between resampled and original average consumption on the 1st platform. The third plot illustrates the between resampled and original average consumption of 2nd platform. Perhaps the posterior predictive checks are reasonable despite poor parameter estimates because the model is over-parameterized. The overestimation of the Gamma parameters is compensated by the bias in the transition matrix parameters.



In summary, the Bernoulli HMM may be more adequate than the Poisson HMM because 1) its parameters have conjugate priors and 2) it can deal with biased frequency data. However, testing the Bernoulli HMM revealed that the binary nature of reach was too weak of a signal to distinguish between different high and low states. To overcome the complications with the hidden Markov models and to more easily incorporate time-varying covariates, we are currently focusing on the third model specification, the Bayesian multivariate logistic regression. We are in the process of testing the model's forecasts and parameter recovery on simulated data. The next steps are to incorporate a better mechanism to capture the anticipated downward trend in digital consumption towards the end of the World Cup. Once the first two weeks of data arrive during the summer, we will put our models to the test and forecast reach for the final three weeks of the championship.

References

[1] Cecile Amblard and Stephane Girard, *A new extension of bivariate fgm copulas*, Metrika **70** (2009), no. 1, 1–17, M3: Article.

[2] Christopher M. Bishop, *Pattern recognition and machine learning*, Springer, New York, 2006, 9780387310732; Christopher M. Bishop.

[3] Peter J. Danaher and Michael S. Smith, *Modeling multivariate distributions using copulas: Applications in marketing*, Marketing Science, forthcoming (2009).

[4] Nigel Meade and Towhidul Islam, *Using copulas to model repeat purchase behaviour: an exploratory analysis via a case study*, European Journal of Operational Research **200** (2010), no. 3, 908–917, M3: Article.

[5] Young-Hoon Park and Peter S. Fader, *Modeling browsing behavior at multiple websites*, Marketing Science **23** (2004), no. 3, 280–303, M3: Article.

[6] Press Release, *Espn launches unprecedented cross-media research initiative: Espn xp*, Tech. report, March 22, 2010.

[7] Wikipedia, *Hidden markov model illustration.*